IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
, Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

Implementation and Analysis of Machine Learning Models for Crop Yield Prediction and Fertilizer Recommendation

Mrs. Ashvini Ganapti Patil[1], Dr. Rahul J. Jadhav[2]
[1]Research Student, Bharati Vidyapeeth (Deemed to be University),Yashwantro Mohite Institute of Management, Karad, India
ashvini.patil18@gmail.com
[2]Associate Professor, Bharati Vidyapeeth (Deemed to be University), Institute of Mgmt. and Entrepreneurship Development, Pune, Maharashtra
rahul.jadhav@bharatividyapeeth.edu

Abstract

This paper presents the implementation and analysis of a machine learning-based framework designed to predict crop yield and recommend integrated fertilizer applications using extensive soil, water, crop, and weather data. Leveraging historical datasets, the study applies data pre-processing steps and clustering techniques to identify meaningful patterns relevant to fertilizer and crop recommendation. Various machine learning algorithms including decision trees, random forests, support vector machines, and neural networks are deployed and evaluated. The results demonstrate significant improvements in prediction accuracy and provide actionable insights for optimizing agricultural strategies. This integrated approach supports sustainable farming by enabling precise resource allocation, reducing fertilizer misuse, and enhancing crop productivity, thereby contributing to food security and environmental conservation.

Keywords: Soil Fertility, Machine Learning, Classification, Random Forest, SVM, Agriculture Data Mining.

I. INTRODUCTION

India’s agriculture holds a critical place in the nation's economy, contributing about 15.96% to the GDP. It supports nearly half the workforce and remains fundamental for food security and rural livelihood. Agricultural productivity is influenced by multiple factors including soil composition, weather variability, precipitation, and the application of fertilizers and pesticides. Soil itself, a complex mixture of minerals, organic matter, and water, plays an essential role. The nutrient makeup of soil directly affects plant growth; deficiencies necessitate detailed soil and water analysis to guide appropriate fertilizer use. Fertilizers help boost natural soil fertility and promote plant development. They range from organic to inorganic forms; organic fertilizers improve soil texture and moisture retention, while inorganic ones can lead to increased soil acidity if misused, calling for balanced, integrated fertilizer management.

A significant challenge is the farmers’ limited awareness of soil and water chemical properties, often causing indiscriminate fertilizer use. Excessive or insufficient application harms crops, reduces yields, and degrades soil health. Scientific fertilizer recommendations based on soil, water, and leaf analyses are therefore vital to optimize yields and maintain soil vitality, ultimately supporting sustainable agriculture and mitigating soil erosion.

ISSN PRINT 2319 1775 Online 2320 7876
, Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

Agricultural yield, the crop output per unit area, depends heavily on soil data including nutrient levels, pH, moisture, organic matter, and texture. These factors shape root growth, water retention, and microbial activity, making soil analysis indispensable. Precision agriculture leverages such data for informed fertilizer application, irrigation, and crop choice, increasing productivity while limiting nutrient depletion and environmental contamination.

Machine learning models have become pivotal in analyzing historical crop, soil, and weather data to improve yield predictions and farming decisions. Algorithms widely used include linear regression, decision trees, random forests, support vector machines, artificial neural networks, and LSTM networks. These tools facilitate data-driven decisions, lessen resource waste, and enhance site-specific farming practices. Machine learning not only forecasts yields but also optimizes fertilizer use, helps plan irrigation schedules, detects crop diseases and pests via image analysis, and advises on seasonal planting and harvesting. These predictive capabilities transition agricultural management from reactive to proactive, improving yield consistency and promoting sustainable use of soil and water resources.

II. RESEARCH GAP

Despite progress in agricultural yield prediction using machine learning, key gaps remain. Most studies treat soil analysis and crop yield prediction separately, lacking integrated fertilizer recommendation systems that combine soil, water quality, weather, and crop-specific data for optimized inputs. Real-time data and IoT sensor integration are underutilized, limiting dynamic and timely decision-making based on up-to-date field conditions. Models often focus on broad trends rather than region-specific predictions tailored to local soil and climate variations, reducing practical relevance.

A significant barrier is the shortage of farmer-friendly decision support systems that translate complex model results into actionable insights through accessible interfaces like mobile apps and dashboards. Data quality issues, including missing values and inconsistent records, further degrade model reliability. Moreover, economic analyses weighing the cost-effectiveness and benefits of AI-driven agriculture are sparse, constraining policymaker and farmer adoption. Addressing these challenges is critical to realizing the full potential of machine learning for sustainable and efficient farming.

III. METHODOLOGY

Fig. 1 presents a conceptual framework that leverages machine learning techniques in conjunction with soil analysis to optimize agricultural yield prediction and strategic planning. The model is designed to promote sustainable and resource-efficient farming methods. Each phase plays a pivotal role in translating raw agricultural data into meaningful insights for informed decision-making.

ISSN PRINT 2319 1775 Online 2320 7876
, Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

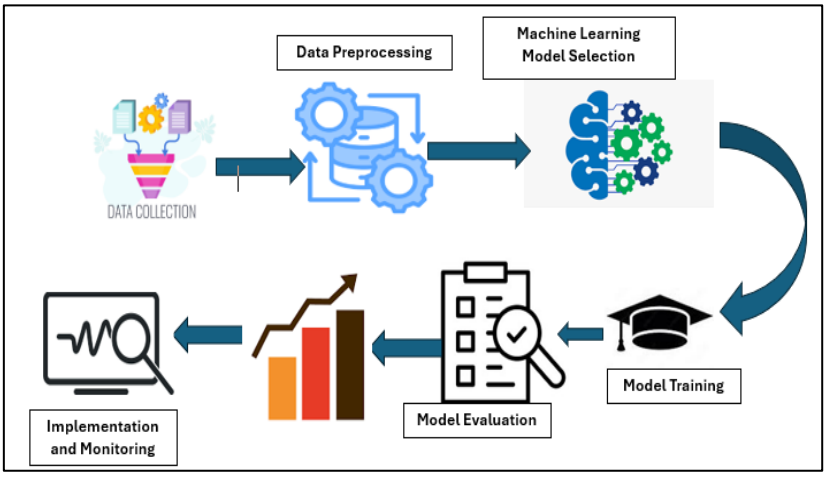


Fig. 1. Proposed Framework Model

Data Collection: Soil samples containing features like N, P, K, pH, and EC. The dataset contains various soil parameters including a categorical target label representing fertility levels. The historical data collection is shown below.

```
#importing libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import warnings
warnings.filterwarnings('ignore')
```

```
#importing the dataset
data = pd.read_csv('f2.csv')
data.head()
```

	Temparature	Humidity	Moisture	Soil_Type	Crop_Type	Nitrogen	Potassium	Phosphorous	Fertilizer
0	20	83	26	Clayey	rice	90	49	36	Urea
1	25	84	32	Loamy	rice	66	59	36	Urea
2	33	64	50	Loamy	Wheat	41	0	0	Urea
3	34	65	54	Loamy	Wheat	38	0	0	Urea
4	38	72	51	Loamy	Wheat	39	0	0	Urea

Fig. 2. Historical Data Collection

Data Preprocessing - The dataset comprises 552 soil and crop samples and provides valuable insights into various environmental and agronomic features critical for soil fertility classification and fertilizer recommendations. It includes 14 distinct fertilizer types such as Urea, TSP, DAP, and a variety of NPK combinations like '28-28' and '10-26-26'. Additionally, the dataset features 20 crop types, including rice, wheat, cotton, maize, and sugarcane, offering a broad representation of agricultural diversity. Numerical data show a wide range in environmental parameters: temperature ranges from 8.8°C to 38°C, humidity from 20% to 95%, and moisture from 0% to 55%. Nutrient values also vary considerably, with nitrogen levels ranging from 0 to 120, potassium from 0 to 100, and phosphorus from 0 to 54, indicating diverse soil fertility conditions. Among soil types, 'Loamy' appears most often, while 'Cotton' is the most frequently grown crop. Urea is the most commonly used fertilizer. This diversity enables robust ML model development for precision agriculture applications.

ISSN PRINT 2319 1775 Online 2320 7876
, Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

```
data['Fertilizer'].unique()
```

```
array(['Urea', 'TSP', 'Superphosphate', 'Potassium sulfate.',
       'Potassium chloride', 'DAP', '28-28', '20-20', '17-17-17',
       '15-15-15', '14-35-14', '14-14-14', '10-26-26', '10-10-10'],
      dtype=object)
```

```
data['Crop_Type'].unique()
```

```
array(['rice', 'Wheat', 'Tobacco', 'Sugarcane', 'Pulses', 'pomegranate',
       'Paddy', 'Oil seeds', 'Millets', 'Maize', 'Ground Nuts', 'Cotton',
       'coffee', 'watermelon', 'Barley', 'kidneybeans', 'orange'],
      dtype=object)
```

```
#statistical parameters
data.describe(include='all')
```

	Temparature	Humidity	Moisture	Soil_Type	Crop_Type	Nitrogen	Potassium	Phosphorous	Fertilizer
count	552.000000	552.000000	552.000000	552	552	552.000000	552.000000	552.000000	552
unique	NaN	NaN	NaN	5	17	NaN	NaN	NaN	14
top	NaN	NaN	NaN	Loamy	Cotton	NaN	NaN	NaN	Urea
freq	NaN	NaN	NaN	192	64	NaN	NaN	NaN	108
mean	28.630435	64.557971	42.840580	NaN	NaN	28.521739	10.144928	21.115942	NaN
std	5.088082	11.880236	11.507275	NaN	NaN	29.121989	13.456956	14.920514	NaN
min	0.000000	50.000000	25.000000	NaN	NaN	0.000000	0.000000	0.000000	NaN
25%	26.000000	54.000000	33.000000	NaN	NaN	10.000000	0.000000	8.000000	NaN
50%	29.000000	62.000000	41.000000	NaN	NaN	15.000000	0.000000	20.000000	NaN
75%	32.000000	68.000000	51.000000	NaN	NaN	37.000000	18.000000	36.000000	NaN
max	38.000000	95.000000	65.000000	NaN	NaN	126.000000	59.000000	54.000000	NaN

Fig. 3. Count Plot for Crop Types in Soil Dataset

The bar chart titled "Count Plot for Crop_Type" visualizes the frequency distribution of various crop types in the dataset, generated using Seaborn in Python. The x-axis displays the crop names, including cotton, sugarcane, pulses, millets, wheat, rice, and others, while the y-axis represents the count of instances for each. Cotton has the highest frequency, exceeding 60 samples, indicating it is the most prominent crop in the dataset. Sugarcane, millets, and pulses also show substantial representation, each with counts between 40 and 55. In contrast, crops like coffee, watermelon, and orange have fewer entries, highlighting a potential imbalance. This uneven distribution can bias ML models and should be addressed during preprocessing to ensure balanced and accurate predictions across all crop types.

```
plt.figure(figsize=(13, 5))
sns.set(style="whitegrid")
sns.countplot(data=data, x='Crop_Type')
plt.title('Count Plot for Crop_Type')
plt.xlabel('Crop_Type')
plt.ylabel('Count')
plt.show()
```

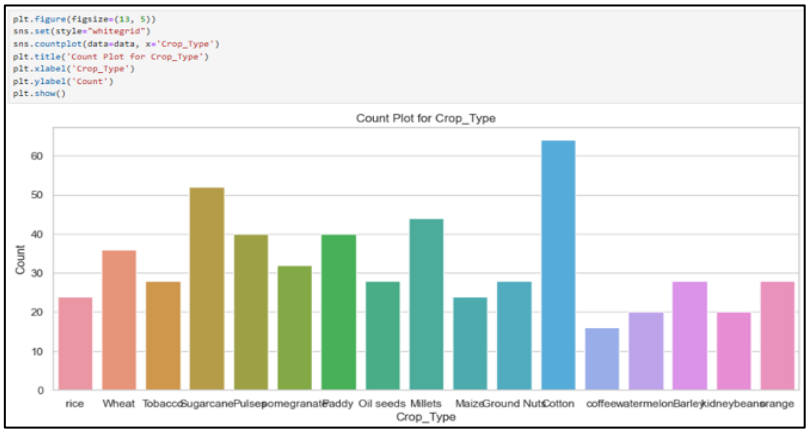


Fig. 4. Count Plot for Crop_Type

IV. IMPLEMENTATION

Once the data is collected, an exploratory data analysis (EDA) is conducted to gain insights into the dataset's structure and characteristics. Python's Matplotlib and Seaborn libraries are commonly used for visualizing data distributions, correlations, and anomalies. EDA helps identify missing values, outliers, and patterns that inform subsequent analysis. Here, hidden patterns in the dataset are identified using clustering analysis, specifically K-means clustering.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

Features relevant to crop recommendation are selected, scaled, and clustered to uncover underlying patterns.

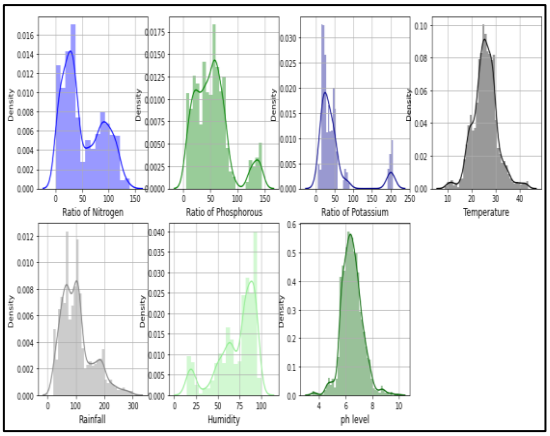


Fig. 5. Distribution for Agricultural Conditions

The above Graph shows us many hidden patterns like many crops need Phosphorous and Potassium at very high level. Need for rain, temperature and ph value vary from crop to crop.

Fig. 6. Suitable Crop for all Seasons

```
# By using dataset, Find the best suitable Crop for Summer Season,Winter Season and Rainy Season
print("Summer Crops")
print(data[(data['temperature'] > 30) & (data['humidity'] > 50)]['label'].unique())
print("----------------------------------------------------------------------------------")
print("Winter Crops")
print(data[(data['temperature'] < 20) & (data['humidity'] > 30)]['label'].unique())
print("----------------------------------------------------------------------------------")
print("Rainy Crops")
print(data[(data['rainfall'] > 200) & (data['humidity'] > 30)]['label'].unique())
```

Python

```
Summer Crops
['pigeonpeas' 'mothbeans' 'blackgram' 'mango' 'grapes' 'orange' 'papaya']
----------------------------------------------------------------------------------
Winter Crops
['maize' 'pigeonpeas' 'lentil' 'pomegranate' 'grapes' 'orange']
----------------------------------------------------------------------------------
Rainy Crops
['rice' 'papaya' 'coconut']
```

In the context of agricultural data, this may include deriving new variables from existing ones or encoding categorical variables. Python libraries like Scikit-learn provide tools for feature scaling, encoding, and selection to prepare the data for modeling. This part trains a predictive model (Random Forest Classifier) to recommend crops based on different features. Features are selected, and the dataset is split into training and testing sets. The model is trained and evaluated for accuracy.

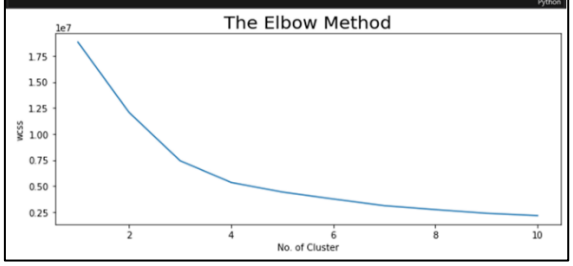
Fig. 7. Use of Elblow Method

ISSN PRINT 2319 1775 Online 2320 7876
, Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

```python
#Determining Optimum number of Clusters within Dataset by using K-means Clustering
plt.rcParams['figure.figsize'] = (10,4)

wcss = []
for i in range(1,11):
    km = KMeans(n_clusters = i,init = 'k-means++',max_iter = 300, n_init = 10, random_state = 0)
    km.fit(x)
    wcss.append(km.inertia_)

#Plotting the Results
plt.plot(range(1,11),wcss)
plt.title('The Elbow Method',fontsize = 20)
plt.xlabel('No. of Cluster')
plt.ylabel('wcss')
plt.show()
```



By using Elbow method, we concluded that dataset consists of four clusters. To implement the Elbow method in Python, we first preprocess the dataset. Once the dataset is prepared, we apply the K-means clustering algorithm with varying numbers of clusters. For each value of k, we compute the sum of squared distances from each data point to its assigned cluster center. This sum, also known as the inertia, is a measure of how tightly the data points are clustered around the centroids.

Fig. 8. Applying K-Means Clustering Analysis

```python
#Implementing K-means Algorithm to perform Clustering Analysis
km = KMeans(n_clusters = 4,init = 'k-means++',max_iter = 300, n_init = 10, random_state = 0)
y_means = km.fit_predict(x)

#Lets find out results
a = data['label']
y_means = pd.DataFrame(y_means)
z = pd.concat([y_means, a],axis = 1)
z = z.rename(columns = {0: 'cluster'})

#Checking Clusters of Each crop
print("Checking results after applying K-means Clustering Analysis \n")
print("Crops in First Cluster:", z[z['cluster'] == 0]['label'].unique())
print("-----------------------------------------------------------")
print("Crops in Second Cluster:", z[z['cluster'] == 1]['label'].unique())
print("-----------------------------------------------------------")
print("Crops in Third Cluster:", z[z['cluster'] == 2]['label'].unique())
print("-----------------------------------------------------------")
print("Crops in Forth Cluster:", z[z['cluster'] == 3]['label'].unique())
```

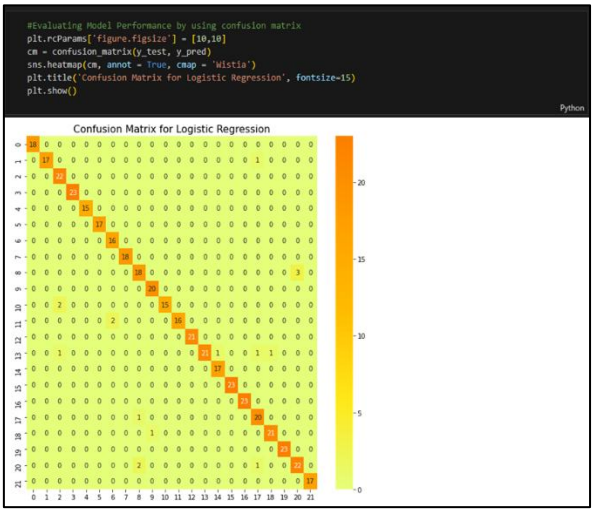
```
Checking results after applying K-means Clustering Analysis

Crops in First Cluster: ['rice' 'pigeonpeas' 'papaya' 'coconut' 'jute' 'coffee']
-----------------------------------------------------------
Crops in Second Cluster: ['maize' 'chickpea' 'kidneybeans' 'pigeonpeas' 'mothbeans' 'mungbean'
 'blackgram' 'lentil' 'pomegranate' 'mango' 'orange' 'papaya' 'coconut']
-----------------------------------------------------------
Crops in Third Cluster: ['grapes' 'apple']
-----------------------------------------------------------
Crops in Forth Cluster: ['maize' 'banana' 'watermelon' 'muskmelon' 'papaya' 'cotton' 'coffee']
```

Next, we plot the values of k against the corresponding inertia scores. The Elbow method suggests selecting the value of k at the "elbow" point of the curve, where the inertia starts to decrease at a slower rate. This point indicates the optimal number of clusters that best capture the underlying structure of the data without overfitting. Deploy various ML algorithms for crop yield prediction, each selected based on its strengths and weaknesses: Decision trees (interpretability, ease of understanding - Ensemble methods like random forests (combining multiple models for improved accuracy) - Support vector machines (handling complex relationships in high-dimensional data) - Neural networks (capturing nonlinear patterns)

Fig. 9. Evulating Model Performance by using Confusion Matrix

ISSN PRINT 2319 1775 Online 2320 7876
 Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023



This section combines all previous steps and executes them in a main function. It ensures that the exploratory analysis, hidden patterns identification, and crop recommendation are performed sequentially. Once the optimal number of clusters is determined, the next step involves training a predictive model, such as a Random Forest Classifier, to recommend crops based on different features. After training the model, it is evaluated using a confusion matrix to assess its performance in terms of accuracy, precision, recall, and F1-score.

V. DISCUSSION

The integration of machine learning with detailed soil and environmental data offers a powerful framework for enhancing sustainable agricultural practices. This approach helps tailor fertilizer recommendations and crop selection to specific soil conditions and climatic factors, significantly reducing resource wastage and environmental harm. The use of clustering methods reveals hidden patterns in soil nutrient requirements and crop suitability across different seasons, supporting more precise and adaptive farming practices. However, real-world implementation challenges persist, including the need for high-quality, comprehensive datasets and the complexity of deploying models in diverse agricultural regions. Addressing these challenges will improve the practical applicability of machine learning models, making them valuable tools for farmers and agricultural policymakers.

VI. CONCLUSION

This study validates the efficacy of machine learning algorithms in accurately predicting crop yields and recommending optimal fertilizers based on multifaceted agricultural data. The integrated model facilitates proactive farm management, reducing overuse of fertilizers and contributing to soil health preservation. Looking ahead, incorporating real-time data from IoT sensors could enhance prediction responsiveness and precision, enabling dynamic adjustments in farming strategies. Future research should emphasize developing region-specific models to account for geographic variability and creating user-friendly decision support systems tailored for smallholder farmers. Additionally, conducting comprehensive economic impact assessments will be essential to encourage wider adoption and policy support for machine learning-driven precision agriculture, thereby fostering sustainable agricultural development and food security.

ISSN PRINT 2319 1775 Online 2320 7876
 Journal UGC CARE Listed (Group-I) Volume 12, Issue , 01 2023

References

1. H.A. Burhan, "Crop Yield Prediction by Integrating Meteorological and Pesticides Use Data with Machine Learning Methods: An Application for Major Crops in Turkey", Journal of Research in Economics, Politics & Finance, 2022, 7(Special Issue): 1-18
2. Gade A. N., Growth of Grapevine Cultivation In Sangli District, Indian Streams Research Journal, ISSN 2230-7850, May-2020
3. A.V. Mhetre, R. V. Chavan, S. A. Chaudhari, Economic Analysis of Grape Production in Sangli District of Maharashtra, International Journal of Current Microbiology and Applied Sciences, ISSN: 2319-7706 Special Issue-11 pp. 1439-1444
4. Khaki S and Wang L (2019) Crop Yield Prediction Using Deep Neural Networks. Front. Plant Sci. 10:621. doi: 10.3389/fpls.2019.00621
5. Russello, H. (2018), Convolutional neural networks for crop yield prediction using satellite images. IBM Center for Advanced Studies.
6. B. K. Punia1, Priyanka Yadav, Predictive Estimates of Employees' Intelligence at Workplace with Special Reference to Emotional and Spiritual Intelligence, BIJIT - BVICAM's International Journal of Information Technology, BIJIT – 2015; January - June 2015; Vol. 7 No. 1; ISSN 0973 – 5658, PP NO- 845-852
7. Marko, O., Brdar, S., Panic, M., Lugonja, P., and Crnojevic, V. (2016). Soybean varieties portfolio optimisation based on yield prediction. Comput. Electron. Agric. 127, 467–474. doi: 10.1016/j.compag.2016. 07.009
8. D.A. Sumner, "Table Grapes Scarlet Royal Mid-Season Maturing Costs & Returns Study," University of California, Department of Agriculture and Natural Resources, 2018.
9. Patel, R., & Gupta, S. (2021), Integrated approaches for crop yield prediction: A review. Journal of Agricultural Informatics, 9(2), 89-104.
10. Wang, L., & Li, H. (2018), Challenges and opportunities in data-driven agriculture: A review. Journal of Agricultural Information Science, 15(1), 35-50.
11. Kumar, S., & Singh, R. (2020), ML techniques for yield prediction in precision agriculture: A review. Computers and Electronics in Agriculture, 184, 106009.
12. Brown, M., & Wilson, D. (2019), Soil properties and their impact on crop performance: A review. Agricultural Research Reviews, 17(3), 128-142.
13. Das, B. S., & Mohanty, B. P. (2006). Root zone soil moisture assessment using remote sensing and unsaturated flow modeling. Journal of Hydrology, 329(1-2), 44-56. https://doi.org/10.1016/j.jhydrol.2006.02.008
14. Espejo-García, B., Mylonas, N., Athanasakos, L., Fountas, S., & Vasilakoglou, I. (2020), Combining generative adversarial networks and deep learning for robust soil moisture prediction. Computers and Electronics in Agriculture, 168, 105135. https://doi.org/10.1016/j.compag.2019.105135
15. Fuentes, S., De Bei, R., Pech, J., & Tyerman, S. (2012), Computational water stress indices obtained from thermal image analysis of grapevine canopies. Irrigation Science, 30(6), 523-536. https://doi.org/10.1007/s00271-012-0388-7